机器学习导论
综合能力测试

151250104, 卢以宁, kiwiloveskiwis@gmail.com

2017 年 6 月 16 日

1 [40pts] Exponential Families

指数分布族 (Exponential Families) 是一类在机器学习和统计中非常常见的分布族, 具有良好的性质。在后文不引起歧义的情况下, 简称为指数族。

指数分布族是一组具有如下形式概率密度函数的分布族群:

$$f_X(x|\theta) = h(x)\exp\left(\eta(\theta)\cdot T(x) - A(\theta)\right) \tag{1.1}$$

其中, $\eta(\theta)$, $A(\theta)$ 以及函数 $T(\cdot)$, $h(\cdot)$ 都是已知的。

(1) **[10pts]** 试证明多项分布 (Multinomial distribution) 属于指数分布族。

(2) **[10pts]** 试证明多元高斯分布 (Multivariate Gaussian distribution) 属于指数分布族。

(3) **[20pts]** 考虑样本集 $\mathcal{D} = \{x_1, \cdots, x_n\}$ 是从某个已知的指数族分布中独立同分布地 (i.i.d.) 采样得到, 即对于 $\forall i \in [1, n]$, 我们有 $f(x_i|\boldsymbol{\theta}) = h(x_i)\exp\left(\boldsymbol{\theta}^{\mathrm{T}}T(x_i) - A(\boldsymbol{\theta})\right)$.

对参数 $\boldsymbol{\theta}$, 假设其服从如下先验分布 :

$$p_\pi(\boldsymbol{\theta}|\boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu)\exp\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\chi} - \nu A(\boldsymbol{\theta})\right) \tag{1.2}$$

其中, $\boldsymbol{\chi}$ 和 ν 是 $\boldsymbol{\theta}$ 生成模型的参数。请计算其后验, 并证明后验与先验具有相同的形式。(**Hint**: 上述又称为"共轭"(Conjugacy), 在贝叶斯建模中经常用到)

Solution. (1)

$$\begin{aligned} f_X(x|\mathbf{p}) &= \frac{n!}{x_1!\cdots x_k!}p_1^{x_1}\times\cdots\times p_k^{x_k} \\ &= \frac{n!}{x_1!\cdots x_k!}\exp(\ln p_1\times x_1\times\cdots\times\ln p_k\times x_k) \end{aligned} \tag{1.3}$$

其中 $n = \sum_i x_i$. 则令 $h(x) = \frac{n!}{x_1!\cdots x_k!}$, $\eta(\mathbf{p}) = \ln\mathbf{p}, T(\mathbf{x}) = \mathbf{x}, A(\mathbf{p}) = 0$ 即可。

(2)

$$
\begin{aligned}
f_X(x|\Sigma,\mu) &= \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}} \\
&= \frac{\exp\left(-\frac{1}{2}(\mathbf{x}^{\mathbf{T}}\Sigma^{-1}\mathbf{x}-2\mathbf{x}^{\mathbf{T}}\Sigma^{-1}\boldsymbol{\mu}+\boldsymbol{\mu}^{T}\Sigma^{-1}\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}} \\
&= \frac{\exp\left(\mathrm{vec}(-\frac{1}{2}\Sigma^{-1})^{T}\mathrm{vec}(\mathbf{x}\mathbf{x}^{\mathbf{T}})+\mathrm{vec}(\Sigma^{-1}\boldsymbol{\mu})^{T}\mathrm{vec}(\mathbf{x}^{\mathbf{T}})-\frac{1}{2}\boldsymbol{\mu}^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\mu}-\frac{1}{2}\ln|\boldsymbol{\Sigma}|\right)}{\sqrt{(2\pi)^k}}
\end{aligned}
\tag{1.4}
$$

则令 $h(x)=\frac{1}{\sqrt{(2\pi)^k}}$, $\eta(\boldsymbol{\Sigma},\boldsymbol{\mu})=\begin{bmatrix}-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\\ \boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\mu}\end{bmatrix}$, $T(\mathbf{x})=\begin{bmatrix}\mathbf{x}\mathbf{x}^{\mathbf{T}}\\ \mathbf{x}^{\mathbf{T}}\end{bmatrix}$, $A(\Sigma,\mu)=\frac{1}{2}\boldsymbol{\mu}^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\mu}+\frac{1}{2}\ln|\boldsymbol{\Sigma}|$ 即可。

(3)

$$
f(\mathbf{X}|\boldsymbol{\theta})=(\prod_{i=1}^{n}h(x_i))\exp\left(\boldsymbol{\theta}^{\mathrm{T}}\sum_{i=1}^{n}T(x_i)-nA(\boldsymbol{\theta})\right) \tag{1.5}
$$

$$
p_\pi(\boldsymbol{\theta}|\boldsymbol{\chi},\nu,\mathbf{X})\propto(\prod_{i=1}^{n}h(x_i))f(\boldsymbol{\chi},\nu)\exp\left(\boldsymbol{\theta}^{\mathrm{T}}(\boldsymbol{\chi}+\sum_{i=1}^{n}T(x_i))-(\nu+n)A(\boldsymbol{\theta})\right) \tag{1.6}
$$

从而, 1.5 与 1.6 形式相同。

2 [40pts] Decision Boundary

考虑二分类问题, 特征空间 $X \in \mathcal{X} = \mathbb{R}^d$, 标记 $Y \in \mathcal{Y} = \{0, 1\}$. 我们对模型做如下生成式假设 :

- attribute conditional independence assumption: 对已知类别, 假设所有属性相互独立, 即每个属性特征独立地对分类结果发生影响 ;
- Bernoulli prior on label: 假设标记满足 Bernoulli 分布先验, 并记 $\Pr(Y = 1) = \pi$.

(1) **[20pts]** 假设 $P(X_i|Y)$ 服从指数族分布, 即

$$\Pr(X_i = x_i|Y = y) = h_i(x_i) \exp(\theta_{iy} \cdot T_i(x_i) - A_i(\theta_{iy}))$$

请计算后验概率分布 $\Pr(Y|X)$ 以及分类边界 $\{x \in \mathcal{X} : P(Y = 1|X = x) = P(Y = 0|X = x)\}$. (**Hint**: 你可以使用 sigmoid 函数 $\mathcal{S}(x) = 1/(1 + e^{-x})$ 进行化简最终的结果).

(2) **[20pts]** 假设 $P(X_i|Y = y)$ 服从高斯分布, 且记均值为 μ_{iy} 以及方差为 σ_i^2 (注意, 这里的方差与标记 Y 是独立的), 请证明分类边界与特征 X 是成线性的。

Solution. (1)

$$\Pr(\mathbf{X}|Y = y) = (\prod_{i=1}^{d} h_i(x_i)) \exp(\sum_{i=1}^{d} (\theta_{iy} \cdot T_i(x_i) - A_i(\theta_{iy}))) \tag{2.1}$$

$$\Pr(\mathbf{X}) = (\prod_{i=1}^{d} h_i(x_i))(\pi \exp(\sum_{i=1}^{d} (\theta_{i1} \cdot T_i(x_i) - A_i(\theta_{i1}))) + (1 - \pi) \exp(\sum_{i=1}^{d} (\theta_{i0} \cdot T_i(x_i) - A_i(\theta_{i0})))) \tag{2.2}$$

$$\begin{aligned} \Pr(Y = 1|\mathbf{X}) &= \frac{\pi}{\left(\pi + (1 - \pi) \exp(\sum_{i=1}^{d} ((\theta_{i0} - \theta_{i1}) \cdot T_i(x_i) + A_i(\theta_{i1}) - A_i(\theta_{i0})))\right)} \\ \Pr(Y = 0|\mathbf{X}) &= \frac{1 - \pi}{\left(1 - \pi + \pi \cdot \exp(\sum_{i=1}^{d} ((\theta_{i1} - \theta_{i0}) \cdot T_i(x_i) + A_i(\theta_{i0}) - A_i(\theta_{i1})))\right)} \end{aligned} \tag{2.3}$$

分类边界:

$$\begin{aligned} &\pi^2 \cdot \exp(\sum_{i=1}^{d} ((\theta_{i1} - \theta_{i0}) \cdot T_i(x_i) + A_i(\theta_{i0}) - A_i(\theta_{i1}))) \\ &= (1 - \pi)^2 \exp(\sum_{i=1}^{d} ((\theta_{i0} - \theta_{i1}) \cdot T_i(x_i) + A_i(\theta_{i1}) - A_i(\theta_{i0}))) \end{aligned} \tag{2.4}$$

化简可得:

$$\ln \frac{\pi}{1 - \pi} = \sum_{i=1}^{d} ((\theta_{i0} - \theta_{i1}) \cdot T_i(x_i) + A_i(\theta_{i1}) - A_i(\theta_{i0})) \tag{2.5}$$

(2) 因为 $P(X_i|Y=y)$ 服从高斯分布，所以:

$$\begin{aligned} h_i(x_i) &= \frac{1}{\sqrt{2\pi}} \\ \theta_{iy} &= \begin{bmatrix} \frac{\mu_{iy}}{\sigma_i^2} \\ \frac{1}{-2\sigma_i^2} \end{bmatrix} \\ T_i(x_i) &= \begin{bmatrix} x_i \\ x_i^2 \end{bmatrix} \\ A_i(\theta_{iy}) &= \frac{\mu_{iy}^2}{2\sigma_i^2} + \ln \sigma_i \end{aligned} \tag{2.6}$$

从而分类边界为:

$$\ln \frac{\pi}{1-\pi} = \sum_{i=1}^{d} ((\frac{\mu_{i0} - \mu_{i1}}{\sigma_i^2}) x_i + A_i(\theta_{i1}) - A_i(\theta_{i0})) \tag{2.7}$$

可见与 x_i 呈线性。

3 [70pts] Theoretical Analysis of k-means Algorithm

给定样本集 $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, k-means 聚类算法希望获得簇划分 $\mathcal{C} = \{C_1, C_2, \cdots, C_k\}$, 使得最小化欧式距离

$$J(\gamma, \mu_1, \ldots, \mu_k) = \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij} ||\mathbf{x}_i - \mu_j||^2 \tag{3.1}$$

其中, $\mu_1, \ldots, \mu_k$ 为 k 个簇的中心 (means), $\gamma \in \mathbb{R}^{n \times k}$ 为指示矩阵 (indicator matrix) 定义如下：若 $\mathbf{x}_i$ 属于第 j 个簇, 则 $\gamma_{ij} = 1$, 否则为 0.

则最经典的 k-means 聚类算法流程如算法1中所示 (与课本中描述稍有差别, 但实际上是等价的)。

Algorithm 1: k-means Algorithm

1 Initialize $\mu_1, \ldots, \mu_k$.

2 **repeat**

3 **Step 1**: Decide the class memberships of $\{\mathbf{x}_i\}_{i=1}^n$ by assigning each of them to its nearest cluster center.

$$\gamma_{ij} = \begin{cases} 1, & ||\mathbf{x}_i - \mu_j||^2 \leq ||\mathbf{x}_i - \mu_{j'}||^2, \forall j' \\ 0, & \text{otherwise} \end{cases}$$

4 **Step 2**: For each $j \in \{1, \cdots, k\}$, recompute μ_j using the updated γ to be the center of mass of all points in C_j:

$$\mu_j = \frac{\sum_{i=1}^n \gamma_{ij} \mathbf{x}_i}{\sum_{i=1}^n \gamma_{ij}}$$

5 **until** *the objective function J no longer changes*;

(1) **[10pts]** 试证明, 在算法1中, **Step 1** 和 **Step 2** 都会使目标函数 J 的值降低.

(2) **[10pts]** 试证明, 算法1会在有限步内停止。

(3) **[10pts]** 试证明, 目标函数 J 的最小值是关于 k 的非增函数, 其中 k 是聚类簇的数目。

(4) **[20pts]** 记 $\hat{\mathbf{x}}$ 为 n 个样本的中心点, 定义如下变量,

total deviation	$T(X) = \sum_{i=1}^n \|\mathbf{x}_i - \hat{\mathbf{x}}\|^2 / n$
intra-cluster deviation	$W_j(X) = \sum_{i=1}^n \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2 / \sum_{i=1}^n \gamma_{ij}$
inter-cluster deviation	$B(X) = \sum_{j=1}^k \frac{\sum_{i=1}^n \gamma_{ij}}{n} \|\mu_j - \hat{\mathbf{x}}\|^2$

试探究以上三个变量之间有什么样的等式关系？基于此, 请证明, k-means 聚类算法可以认为是在最小化 intra-cluster deviation 的加权平均, 同时近似最大化 inter-cluster deviation.

(5) **[20pts]** 在公式(3.1)中, 我们使用 ℓ_2-范数来度量距离 (即欧式距离), 下面我们考虑使用 ℓ_1-范数来度量距离

$$J'(\gamma, \mu_1, \ldots, \mu_k) = \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij} ||\mathbf{x}_i - \mu_j||_1 \tag{3.2}$$

- **[10pts]** 请仿效算法1(k-means-ℓ_2 算法), 给出新的算法 (命名为 k-means-ℓ_1 算法) 以优化公式3.2中的目标函数 J'.
- **[10pts]** 当样本集中存在少量异常点 (outliers) 时, 上述的 k-means-ℓ_2 和 k-means-ℓ_1 算法, 我们应该采用哪种算法？即, 哪个算法具有更好的鲁棒性？请说明理由。

Solution. (1) 在 **Step 1** 中, $\forall i$, 令

$$\hat{j} = \min_{1 \le j \le k} ||\mathbf{x}_i - \mu_j||^2$$

又因为 $\boldsymbol{\gamma}_i$ 是指示向量,

$$\sum_{j=1}^{k} \gamma_{ij} ||\mathbf{x}_i - \mu_j||^2 \ge ||\mathbf{x}_i - \mu_{\hat{j}}||^2$$

故 $J(\gamma, \mu_1, \ldots, \mu_k)$ 若在第一步发生改变, 一定下降。

在 **Step 2** 中, 令 $X_{\in j} = \{\mathbf{x}_i | \gamma_{ij} = 1\}$ 代表在簇 j 中点的集合, $n_j = |X_{\in j}|$ 为该集合大小, $\bar{\mathbf{x}}$ 为该集合的均值。则 $\forall j, \forall \mathbf{a}$:

$$\begin{aligned} \sum_{x \in X_{\in j}} ||\mathbf{x} - \mathbf{a}||^2 &= \sum_{x \in X_{\in j}} (\mathbf{x^T x} + \mathbf{a^T a} - \mathbf{2x^T a}) \\ &= \sum_{x \in X_{\in j}} \mathbf{x^T x} + n_j \mathbf{a^T a} - 2 n_j \bar{\mathbf{x}}^T \mathbf{a} \end{aligned} \tag{3.3}$$

而当 $\mathbf{a} = \bar{\mathbf{x}}$ 时上式取得最小值。故 $J(\gamma, \mu_1, \ldots, \mu_k)$ 若在第二步发生改变, 一定下降。

(2) 由于不同的 $J(\gamma, \mu_1, \ldots, \mu_k)$ 对应不同的 $\boldsymbol{\gamma}$, 且每次更新时 J 均下降 (不下降时终止), 故 $\boldsymbol{\gamma}$ 不会与先前重复。又由于 $\boldsymbol{\gamma}$ 最多有 k^n 种可能取值, 所以算法会在有限步终止。

(3) 令 $k = k_0$ 时取得 J 的最小值的指示矩阵 γ 不变, $k = k_1$ 时将 μ_{k+1} 设为 $\mathcal{D}$ 中任意一个点, 则 J 的值必不上升。从而 J 的最小值不上升。

(4) 令 $X_{\in j} = \{\mathbf{x}_i | \gamma_{ij} = 1\}$ 代表在簇 j 中点的集合, $n_j = \sum_{i=1}^{n} \gamma_{ij}$ 为该集合大小, $\hat{\mathbf{x}}$ 为该集合的均值. 则:

$$\begin{aligned} n_j W_j(X) + n B_j(X) &= \sum_{x \in X_{\in j}} \|\mathbf{x} - \mu_j\|^2 + n_j \|\mu_j - \hat{\mathbf{x}}\|^2 \\ &= \sum_{x \in X_{\in j}} \mathbf{x^T x} - n_j ||\boldsymbol{\mu}_j||^2 + n_j (||\boldsymbol{\mu}_j||^2 + ||\hat{\mathbf{x}}||^2 - 2\boldsymbol{\mu}_j^T \hat{\mathbf{x}}) \\ &= \sum_{x \in X_{\in j}} \mathbf{x^T x} + n_j ||\hat{\mathbf{x}}||^2 - 2 n_i \boldsymbol{\mu}_j^T \hat{\mathbf{x}} \\ &= \sum_{x \in X_{\in j}} \|\mathbf{x} - \hat{\mathbf{x}}\|^2 \end{aligned} \tag{3.4}$$

从而:

$$\sum_{j=1}^{k} \frac{n_j}{n} W_j(X) + B(X) = T(X) \tag{3.5}$$

由于算法迭代过程中 $T(X)$ 不变，而目标函数 J 的值下降, 即 $\sum_{j=1}^{k} \frac{n_j}{n} W_j(X)$ 的值下降, 所以 $B(X)$ 上升。所以可以认为"是在最小化 intra-cluster deviation 的加权平均, 同时近似最大化 inter-cluster deviation"。

(5) 设 d 代表维度。当样本集中存在少数异常点时, k-means-ℓ_1 算法具有更好的鲁棒性。因

Algorithm 2: k-means-ℓ_1 Algorithm

Initialize $\mu_1, \ldots, \mu_k$.

repeat

Step 1: Decide the class memberships:

$$\gamma_{ij} = \begin{cases} 1, & ||\mathbf{x}_i - \mu_j||_1 \leq ||\mathbf{x}_i - \mu_{j'}||_1, \forall j' \\ 0, & \text{otherwise} \end{cases}$$

Step 2: For each $j \in \{1, \cdots, k\}$, recompute μ_j using the updated γ :

$$\forall d, \mu_j[d] = \text{median of } \{\mathbf{x_i} | \gamma_i j = 1\}$$

until *the objective function J no longer changes*;

为采用平均时, 与这些异常点最近的簇的中心点很可能受到较大影响从而偏离本应在的中心, 而采用中位数时，该中心点不会发生太大变化。

4 [50pts] Kernel, Optimization and Learning

给定样本集 $\mathcal{D}=\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\cdots,(\mathbf{x}_m,y_m)\}$, $\mathcal{F}=\{\Phi_1\cdots,\mathbf{\Phi}_d\}$ 为非线性映射族。考虑如下的优化问题

$$\min_{\mathbf{w},\mu\in\Delta_q}\quad \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\|\mathbf{w}_k\|_2^2+C\sum_{i=1}^{m}\max\left\{0,1-y_i\left(\sum_{k=1}^{d}\mathbf{w}_k\cdot\mathbf{\Phi}_k(\mathbf{x}_i)\right)\right\} \tag{4.1}$$

其中, $\Delta_q=\{\boldsymbol{\mu}|\mu_k\geq 0,k=1,\cdots,d;\|\boldsymbol{\mu}\|_q=1\}$.

(1) **[30pts]** 请证明, 下面的问题4.2是优化问题4.1的对偶问题。

$$\begin{aligned}\max_{\boldsymbol{\alpha}}\quad & 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}-\left\|\begin{matrix}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_1\mathbf{Y}\boldsymbol{\alpha}\\ \vdots\\ \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_d\mathbf{Y}\boldsymbol{\alpha}\end{matrix}\right\|_p\\ \text{s.t.}\quad & \mathbf{0}\leq\boldsymbol{\alpha}\leq\mathbf{C}\end{aligned} \tag{4.2}$$

其中, p 和 q 满足共轭关系, 即 $\frac{1}{p}+\frac{1}{q}=1$. 同时, $\mathbf{Y}=\mathrm{diag}([y_1,\cdots,y_m])$, $\mathbf{K}_k$ 是由 $\mathbf{\Phi}_k$ 定义的核函数 (kernel).

(2) **[20pts]** 考虑在优化问题4.2中, 当 $p=1$ 时, 试化简该问题。

Solution. 一个偷懒的人在此悄悄回望.